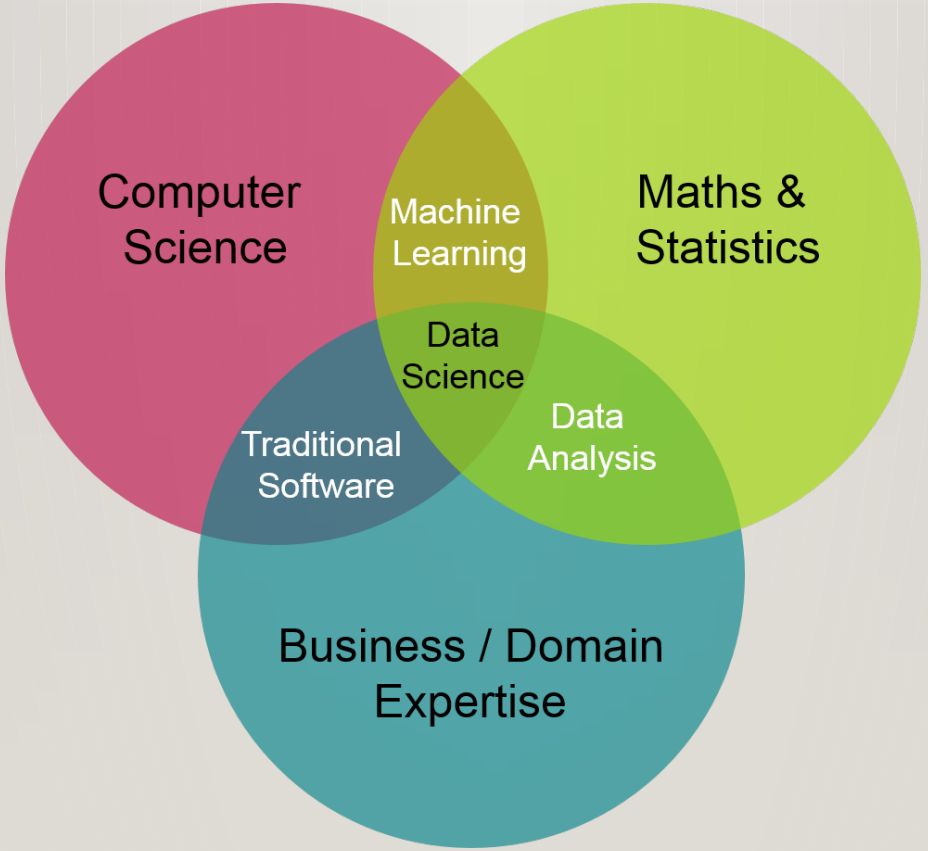
MAKING A CAREER IN DATA SCIENCE: A BEGINNER'S GUIDE

MUHAMMAD HAMMAD KHAN DATA SCIENTIST AT DARAZ @ ALIBABA GROUP

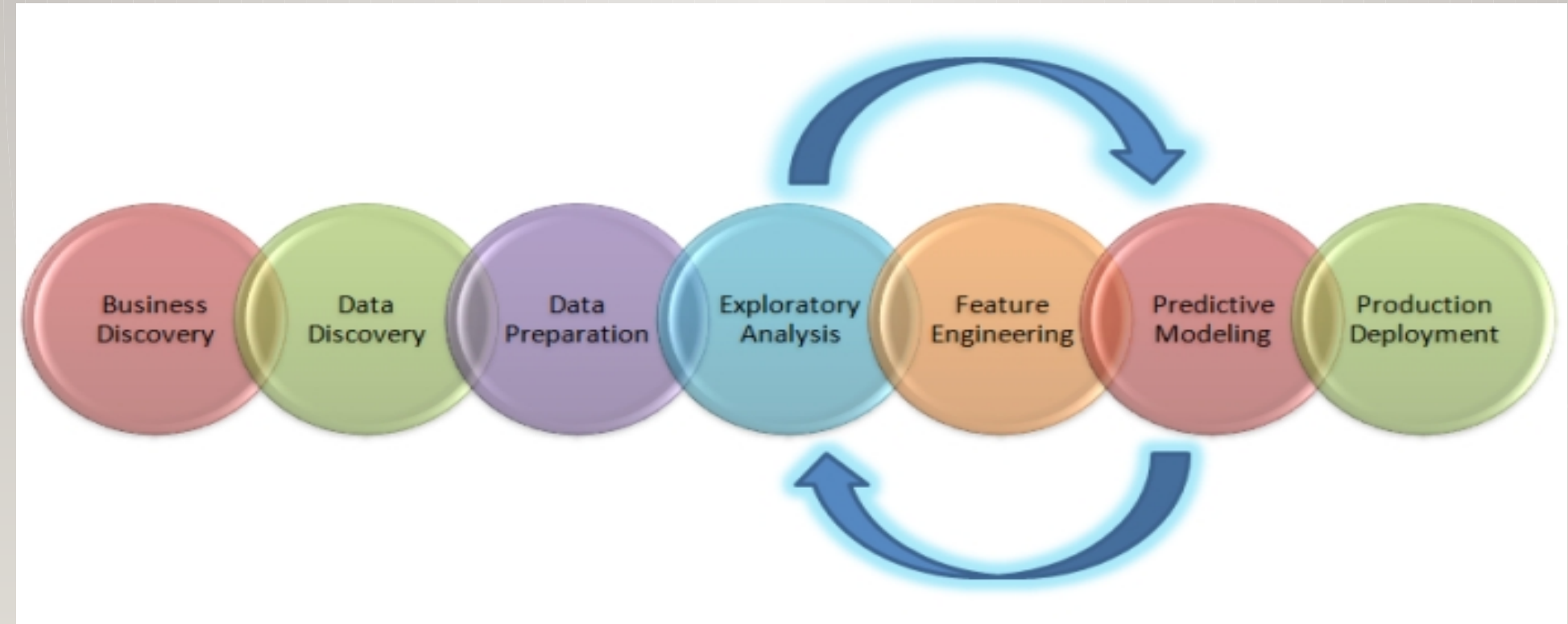
TOPICS WE WILL BE COVERING.

- What is data science?
- Data science project lifecycle.
- Different roles in data science ecosystem.
- Exploring your passion in data science.
- What is Kaggle.
- Learning roadmap of data science.
- Importance of math in data science.
- Key lessons to be successful.
- Common mistakes / Myths aspiring data scientist makes.
- How to get your first Job.

WHAT IS DATA SCIENCE?



DATA SCIENCE PROJECT LIFECYCLE



DIFFERENT ROLES IN DATA SCIENCE ECOSYSTEM:

Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:

SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:

Hadoop, NoSQL, and Python

Data Analysts

also known as business Analysts.

They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

Will use programmes such as:

Excel, Tableau, SQL

WHY DO YOU WANT TO BE A DATA SCIENTIST?

- **Why**—I want to wake up every day excited to go to work so my life would be more fulfilling.
- **How**—Find a career in a field that fascinates me, work with driven and interested people on interesting challenges, and work on the cutting edge of that field.
- **What**—I happen to be a Data Scientist

DO YOU REALLY WANT TO BE A DATA SCIENTIST?

- Are you willing to spend two weeks cleaning data from ten sources, feature engineering for another week, and then testing your model to realize it is worthless?
- Are you willing to understand the nuts and bolts—the math, the concepts, how optimization works?
- Are you going to spend your time studying early in the morning and late at night while your friends go to the concerts/mall and progress through their high paying engineering careers?
- If not, then Data Science (probably) is not for you. I say probably because there are always exceptions.

DO YOU REALLY WANT TO BE A DATA SCIENTIST?

- As a field, Data Science is rapidly changing. The deeper you go, the harder it gets to keep up.
- You need to be committed to learning and studying, and still willing to go back to those first principles and wrestle with the data.
- You need a dizzying array of technical, analytical, and personal skills.
- The day you stop studying is the day you start falling into obsolescence.
- Imagine how exhausting that could be if you do not love what you are doing.

WHICH ELECTIVE COURSES TO TAKE DURING UNDERGRAD STUDIES.

- Information Retrieval and Text Mining.
- Data Science.
- Computer Vision.
- Deep Learning.
- Data Mining.
- Data Warehousing.

ROADMAP TO LEARN DATA SCIENCE.

- Math
  - Linear Algebra.
  - Calculus.
  - Statistics.
  - Probability.
- Programming
  - Python or R.
- SQL.
- Machine Learning.

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing package e.g. R
- Databases SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

BOTTOM UP VS TOP DOWN LEARNING APPROACH.

- Bottom up learning
  - Focused on deep mathematical knowledge.
- Top down learning
  - Focused on practical application more.
- Hybrid approach
  - Balanced approach essential to learn data science.

IMPORTANCE OF MATH IN DATA SCIENCE

- There are many reasons why the mathematics of Machine Learning is important and I will highlight some of them below:
  - Selecting the right algorithm which includes giving considerations to accuracy, training time, model complexity, number of parameters and number of features.
  - Choosing parameter settings and validation strategies.
  - Identifying underfitting and overfitting by understanding the Bias-Variance tradeoff.
  - Estimating the right confidence interval and uncertainty.

- **Kaggle** is an online community of data scientists and machine learners, owned by Google, Inc.
- Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.
- Kaggle got its start by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and short form AI education

DATA SCIENCE ON KAGGLE VS INDUSTRY

- Kaggle
  - Make a **perfect** model with as much time as you have
  - Success is a **high CV score**
  - **Interpretability has no importance**
  - Clean data given in csv format mostly
  - Well defined problem statements
  - No model deployment

DATA SCIENCE ON KAGGLE VS INDUSTRY

- *Data Science Industry*
  - Make an **impactful** model considering time constraints and other priorities
  - Success is **improving product or business** (or team!)
  - **Interpretability helps for persuasion**
  - Dirty and a lot of missing data.
  - No well defined problem statements
  - Model deployment and testing.
  - Story telling skills required

CONTINUOUS LEARNING IS IMPORTANT IN DATA SCIENCE. HOW TO LEARN ON REGULAR BASIS?

- Kaggle Kernels.
- Medium Blogs.
- YouTube channels.
- LinkedIn.
- Mentor.
- MOOC’s

THERE ARE 8 KEY LESSONS THAT I ATTRIBUTE MY SUCCESS TO:

- Pick ONE programming language and STICK to it (Python or R).
- Be clear about your motivation. Learning Data Science is HARD and time consuming So it’s easy to lose motivation when on the journey.
- Immerse yourself in the community. You need to surround yourself with all things Data Science. (Blogs, Subscribing YouTube channels, FB groups Etc.)

THERE ARE 8 KEY LESSONS THAT I ATTRIBUTE MY SUCCESS TO:

- Don’t get stuck in the Tutorial carousel. If you keep doing tutorial after tutorial, it’s easy to deceive yourself into thinking you know what you’re doing
- Pick a small set of resources. There are SO many resources out there to learn the fundamentals of Data Science. Make a curriculum and stick to it.
- Go to Hackathons! Don’t wait till you’re “ready” before you go to a hackathon. (Kaggle)

THERE ARE 8 KEY LESSONS THAT I ATTRIBUTE MY SUCCESS TO:

- Find a mentor. My mentors ended up being influential Data Scientists who I interacted with by following them on LinkedIn.(Seniors or people from the industry)

- Be prepared to sacrifice your weeknights and weekends. You have to put in a lot of deliberate practice and spend significant time studying, expect your social life to suffer. (Work hard consistently)

COMMON-MISTAKES-ASPIRING-FRESHER-DATA-SCIENTISTS-MAKE

- Learning Theoretical Concepts without Applying Them
- Heading Straight for Machine Learning Techniques without Learning the Prerequisites
- Relying Solely on Certifications and Degrees
- Assuming that what you see in ML Competitions is what Real-Life Jobs are Like
- Focusing on Model Accuracy over Applicability and Interpretability in the Domain
- Using too many Data Science Terms in your Resume
- Giving Tools and Libraries Precedence over the Business Problem

COMMON-MISTAKES-ASPIRING-FRESHER-DATA-SCIENTISTS-MAKE

- Not Spending Enough Time on Exploring and Visualizing the Data (Curiosity)
- Not Having a Structured Approach to Problem Solving
- Trying to Learn Multiple Tools at Once
- Not Studying in a Consistent Manner
- Shying Away from Discussions and Competitions
- Not working on Communication Skills

BREAKING COMMON MYTHS IN DATA SCIENCE

- Career-Related Myths:
  - Ph.D. is Mandatory to Become a Data Scientist
    - Applied Data Science Role
    - Research Role
- Learning a Tool is Enough to Become a Data Scientist:
  - Data science requires a combination of multiple skills. Programming is not at the center of the data science spectrum – it is just one part of a whole.
    - Technical qualities.
    - Non-technical qualities or soft skills.

BREAKING COMMON MYTHS IN DATA SCIENCE

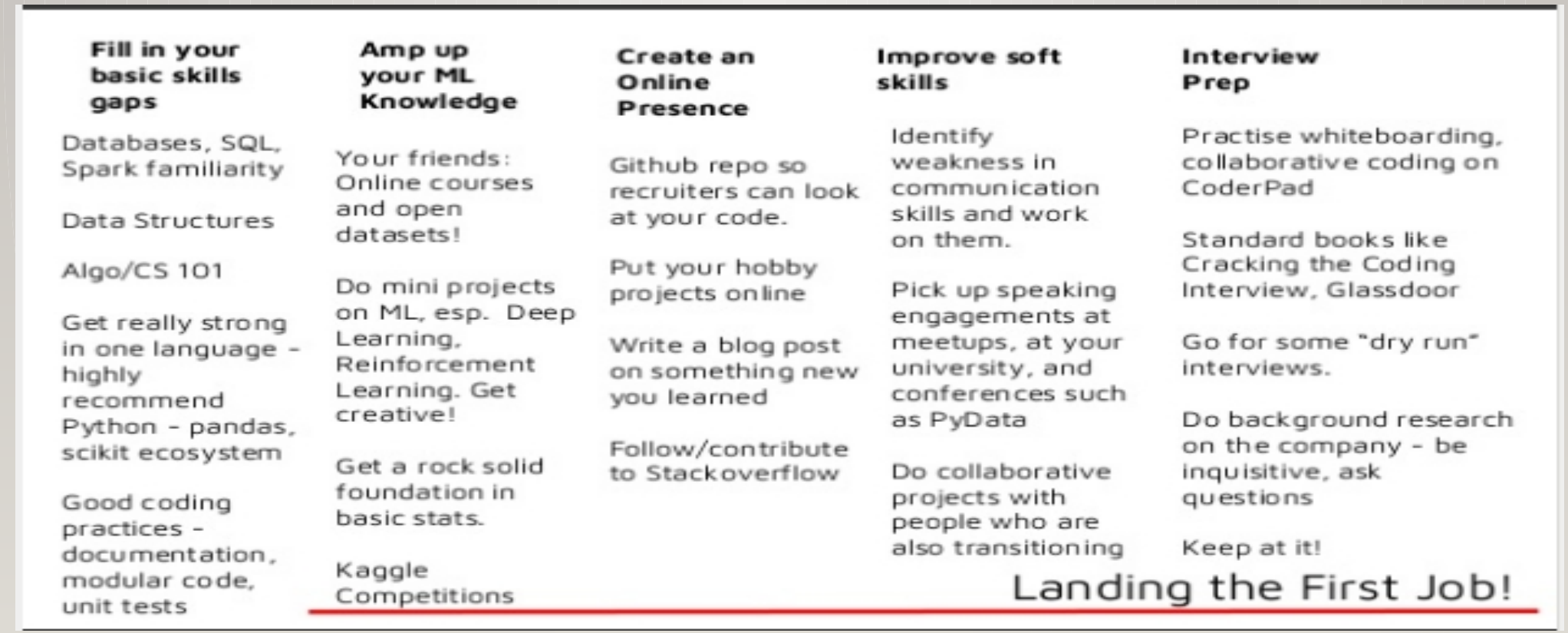
- Data Science is Only About Building Predictive Models:
  - Understanding the problem statement
  - Hypothesis building
  - Data collection
  - Verifying the data
  - Data cleaning
  - Exploratory analysis
  - Designing the model
  - Testing/Verifying the model
    - If an error is found, head back to the verification or cleaning stage
  - Putting it into production (deploying the model)

BREAKING COMMON MYTHS IN DATA SCIENCE

- Data Collection is a Breeze, the Focus should be on Building Models:
  - ***Interviewer:*** *What's your favorite part of a data science role apart from designing models?*
  - ***Fresher DS:*** *I like the feature engineering part.*
  - ***Int:*** *Sounds fair. How do you usually collect data for your projects?*
  - ***Fresher DS:*** *Um, I usually just download it from one of the open-source platforms.*
  - ***Int :*** *OK. but what if the data is skewed or biased? How do you verify the identity of the data? And what will you do when you're asked to collect data from multiple sources that require database skills?*
  - ***Fresher DS:*** *I hadn't thought about that..*

FINDING YOUR FIRST DATA SCIENCE JOB

FINDING YOUR FIRST DATA SCIENCE JOB



FINDING YOUR FIRST DATA SCIENCE JOB

- **Types of Portfolio Projects**
  - Different projects should demonstrate your different skills and abilities.
  - A **Data Cleaning Project** aims to demonstrate that you are capable of taking several noisy datasets, clean them, improve their quality and/or size and combine them to solve a specific problem.
  - A **Data Storytelling Project** should demonstrate that you can get insights from data, communicated them clearly and keep your reader engaged.
  - A **Data Visualization Project** should prove that you are capable of visualizing data with an appropriate choice of plots and charts.

FINDING YOUR FIRST DATA SCIENCE JOB

- A **Machine Learning Project** aims to demonstrate that you are able to build statistical models using arbitrary data, save and load models model, as well as make predictions using the model.
- An **End to End Project** — a project that proves you are capable of building a stand-alone system that can be used in production.
- An **Explanatory Post** in a blog. This post should demonstrate your ability to clearly communicate complex machine learning or statistics concepts to various auditory.

DATA SCIENCE AT DARAZ

- Forecasting demand for goods and services
- Optimizing pricing structures
- Customer/Seller churn.
- Market basket analysis
- Customer sentiment analysis
- Fraud Detection

ANY QUESTIONS?

- You can always ping me for your queries.
- LinkedIn : https://www.linkedin.com/in/hammad-khan-b84822142/
- Email : hammadkhann34@gmail.com

RESOURCES/ REFERENCES

- https://www.analyticsvidhya.com/blog/2015/11/lifetime-lessons-20-data-scientist-today/?utm_source=linkedin.com&utm_medium=social
- https://www.analyticsvidhya.com/blog/2014/11/tips-prepare-cv-data-science-roles/
- https://blog.usejournal.com/how-to-become-a-data-scientist-in-12-months-71aa9ee822d9
- http://www.bigcloud.io/why-you-need-to-keep-learning-in-data-science/
- https://towardsdatascience.com/the-mathematics-of-machine-learning-894f046c568
- https://machinelearningmastery.com/youre-wrong-machine-learning-not-hard/

RESOURCES/ REFERENCES

- https://semanti.ca/blog/?how-to-get-your-first-data-science-job
- https://www.analyticsvidhya.com/blog/2018/07/13-common-mistakes-aspiring-fresher-data-scientists-make-how-to-avoid-them/
- https://www.analyticsvidhya.com/blog/2015/12/job-roles-data-science-industry-who-what/
- https://towardsdatascience.com/the-cold-start-problem-how-to-build-your-machine-learning-portfolio-6718b4ae83e9
- https://www.analyticsvidhya.com/blog/2018/06/comprehensive-data-science-machine-learning-interview-guide/

RESOURCES/ REFERENCES

- https://medium.com/s/story/essential-math-for-data-science-why-and-how-e88271367fbd
- https://towardsdatascience.com/best-resources-for-ai-machine-learning-data-science-d72625d4689d
- https://medium.freecodecamp.org/i-ranked-all-the-best-data-science-intro-courses-based-on-thousands-of-data-points-db5dc7e3eb8e
- https://www.analyticsvidhya.com/blog/2019/01/myths-data-science-transition/?utm_source=linkedin.com&utm_medium=social

RESOURCES/ REFERENCES

- https://medium.freecodecamp.org/if-you-want-to-learn-data-science-take-a-few-of-these-statistics-classes-9bbabab098b9